2017

Cross-Modality Feature Learning Through Generic Hierarchical Hyperlingual-Words

Shao, Ming

Cross-Modality Feature Learning Through Generic Hierarchical Hyperlingual-Words

Ming Shao, *Student Member, IEEE*, and Yun Fu, *Senior Member, IEEE*

***Abstract*— Recognizing facial images captured under visible light has long been discussed in the past decades. However, there are many impact factors that hinder its successful application in real-world, e.g., illumination, pose variations. Recent work has concentrated on different spectrals, i.e., near infrared, that can only be perceived by specifically designed device to avoid the illumination problem. However, this inevitably introduces a new problem, namely, cross-modality classification. In brief, images registered in the system are in one modality, while images that captured momentarily used as the tests are in another modality. In addition, there could be many within-modality variations—pose and expression—leading to a more complicated problem for the researchers. To address this problem, we propose a novel framework called hierarchical hyperlingual-words (Hwords) in this paper. First, we design a novel structure, called generic Hwords, to capture the high-level semantics across different modalities and within each modality in weakly supervised fashion, meaning only modality pair and variations information are needed in the training. Second, to improve the discriminative power of Hwords, we propose a novel distance metric through the hierarchical structure of Hwords. Extensive experiments on multimodality face databases demonstrate the superiority of our method compared with the state-of-the-art works on face recognition tasks subject to pose and expression variations.**

***Index Terms*— Cross-modality face recognition, hyperlingual-words (Hwords), near infrared (NIR), weighted distance metric.**

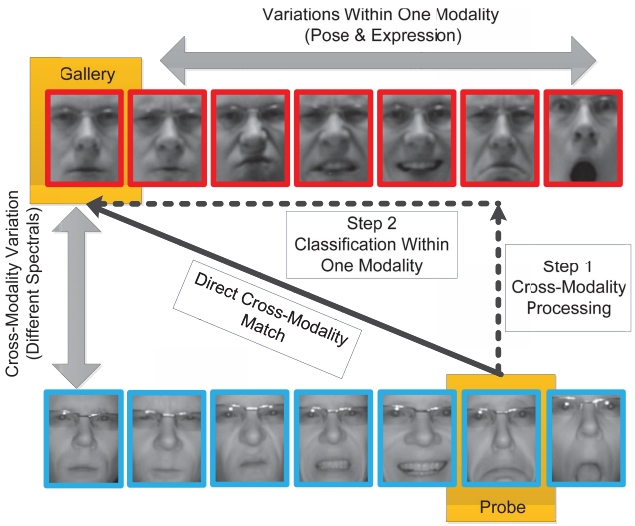

Fig. 1. Illustration of multimodality heterogeneous face recognition. Images in blue border are under NIR, while that in red under VIS. For each subject under a specific spectral, there are seven different expressions: neutral, anger, disgust, fear, happiness, sadness, and surprise, from left to right.

I. Introduction

ILLUMINATION or lighting condition has been identified as one of the most significant impact factors in face recognition [1], [2]. Recently, Li *et al.* [3] exploit near infrared (NIR) images as complements for illumination-free recognition, and the performance turns to be very impressive. However, this approach introduces a new problem that enrolled and test images are in different modalities—while test images are now captured under NIR, huge amount of previously enrolled images are under visible light (VIS). Such cross-modality recognition is indeed challenging, since it is always mixed with other impact factors, e.g., pose and expression.

In fact, such cross-modality problems are not rarely seen, for example, automatic comparisons between the witness's description-based sketch and the mugshot photo in the criminal probe [4]–[6]. The former only includes partial appearance information of the suspect and is drawn by experienced artists, while the latter is the real face photo that objectively renders what a suspect looks like. Remarkable difference between these two modalities makes the direct face match very difficult. Other related examples include synthesizing oil painting or sketch from photos [7], [8]. In a nutshell, the multimodality images discussed in this paper can be summarized as: a group of images of fixed objects captured by different devices and sensors, or recorded in different ways.

In addition to modality variations, both within-class and between-class variations affect the performance of recognition tasks. Such problem was identified as heterogeneous recognition in biometric community and has already been extensively discussed for years [4]–[6], [9]–[17]. We illustrate this problem in Fig. 1 by taking NIR-VIS cross-modality recognition as an example. The two dimensions in Fig. 1 indicate the variations of modality and pose/expression, respectively, and their joint

Manuscript received September 28, 2014; revised December 29, 2015; accepted December 31, 2015. Date of publication January 26, 2016; date of current version January 17, 2017. This work was supported in part by the U.S. Army Research Office Young Investigator under Award W911NF-14-1-0218, in part by the Naval Postgraduate School under Award N00244-15-1-0041, in part by the National Science Foundation through the Division of Computer and Network Systems under Award 1314484, in part by the Office of Naval Research under Award N00014-12-1-1028, and in part by the Office of Naval Research Young Investigator Program under Award N00014-14-1-0484.

M. Shao is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: mingshao@ece.neu.edu).

Y. Fu is with the Department of Electrical and Computer Engineering, College of the Computer and Information Science, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2517014

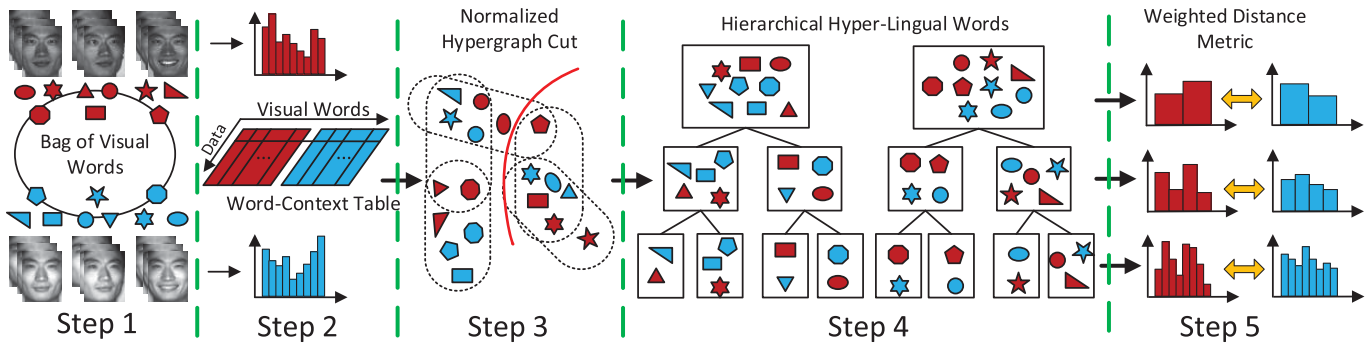

Fig. 2. Framework of the proposed method for cross-modality image classification. Step 1: low-level feature extraction by the BoVW model. Step 2: word-context table construction by training samples. Step 3: hypergraph partition to yield Hwords. Step 4: iterate step 3 to generate H^2words. Step 5: weighted distance metric.

variations are distributed in the 2-D space. A straightforward approach could be: 1) cross-modality processing and 2) classification in the same modality. Following this two-step approach, many state-of-the-art face recognition methods can be directly applied after the cross-modality processing. In the first step, there are mainly three strategies:

1) invariant features [4], [6], [9], [16];
2) common subspace [5], [11], [12], [15], [17];
3) image synthesis [7], [8], [13], [14], [18], [19].

In the second step, we could choose either holistic [20]–[22] or local features [23]–[26] for recognition tasks. Nonetheless, these approaches are not unified in a general framework, and the output of the first step is not necessarily the optimized input for the second step. For example, the cross-modality processing may only concern with the correspondence within a pair of samples, which potentially ruins the discriminant capability that matters in the second step.

Recently, bag of visual words (BoVWs) [27] has been prevalent in computer vision community. If we imagine images in one modality can be written in a specific language (visual words in the codebook), then the former multimodality problem turns to be a translation problem between two languages (visual words in two codebooks). Furthermore, within- and between-class variations can be connected by synonyms and near-synonyms in each language. Similar problems have been widely discussed in machine translation [28] and cross-language document retrieval [29]–[31], where the correlations of the elements in two natural languages are set manually or learned automatically. In cross-modality visual feature learning, however, the correspondence between visual words from different codebooks should be learned, because the semantics of low-level features are usually vague.

In this paper, a novel cross-modality feature learning framework highlighted by generic hyperlingual-words (Hwords) is proposed to address the heterogeneous recognition problem, as shown in Fig. 2. Low-level features, such as visual words, in the BoVW model often suffer from arbitrary variations and lack of high-level semantics, especially on cross-modality problem. Similar to the concept of interlingual used in machine translation [28], the proposed Hwords work exactly as an intermediate language (interlingual) between several sets of codebooks and manage to abridge the semantic gap between mutually exclusive codebooks. In addition, contrary to most existing cross-modality feature learning methods, the proposed generic Hwords can be learned in a weakly supervised fashion. Finally, it is able to incorporate both handcraft and learned codebooks in a flexible way.

There are two tasks included in the proposed generic Hwords: 1) connecting visual words across modalities based on semantics and 2) connecting statistically related words within each modality. If we take these connections as side information, however, the aforementioned problem is already beyond the scope of pairwise graph and its partition. We, therefore, propose to take hypergraph partition to generate Hwords by assuming that the connected visual words are incident with the same hyperedge. The higher order relations encoded in the hyperedge will finally yield Hwords robust to modalities changes as well as other impact factors.

In addition, we further improve the discriminative capability of the proposed Hwords by pyramid histograms match (PM) guided weighted chi-square metric, where we assemble the Hwords in different resolutions into one hierarchical structure, i.e., hierarchical Hwords (H^2words). PM [32], [33] is a principal way to use multiresolution histograms, and it does not rely on the strict local feature correspondence. We utilize the proposed hierarchical structure of Hwords to yield the pyramid histograms, and their matching results lead to a better chi-square metric. In this way, H^2words not only align features from different modalities, but also keep the discriminative power. Comprehensive results on three multimodality face databases demonstrate that the proposed H^2words perform better than the state-of-the-art visual coding scheme and cross-modality feature learning methods, and its hierarchical structure works even better by automatically weighting the metric.

II. Related Work

Invariant features have been explored as the low-level descriptors for cross-modality problems. Difference of Gaussian (DoG) and multiresolution local binary pattern (LBP) were utilized in [9] to match face from NIR to VIS. This thought was further developed in [34], where multi-DoG and multiple features, e.g., histogram of oriented gradients (HOG) [26], gradient location and orientation

histogram (GLOH) [35], and scale-invariant feature transform (SIFT) [25], were integrated by score level fusion. Zhang *et al.* [6] proposed a greedy approach called coupled information-theoretic coding, where an unbalance binary tree was built based on mutual information entropy. Invariant features were extracted from top to bottom by the learned projection in each node for photosketch matching. A novel component-based approach was proposed in [16] to address match problem between composite photos generated by law enforcement agencies and mugshot photos, where a practical system, including facial normalization, partition, feature extraction, and score fusion, was detailed. Recently, learning-based descriptors attracted substantial research attention, where the most discriminant and optimal sampling strategy is learned for local descriptors [36]. More recently, deep face models based on convolutional neural network (CNN) have attracted substantial research attention due to their superior performance on face benchmarks [37], [38]. CNN learns features from local patch through a set of discriminative filters followed by pooling and normalization operations. It then stacks features layer by layer for classification tasks. Notably, autoencoder has been adopted in missing modality problem [39]; however, it considers different settings as it involves auxiliary multimodality data set to help the recognition tasks within single modality.

On the other hand, researchers optimized the common subspace by coupled projections, where cross-modality features remain discriminative. In [10], a common feature space was derived by considering both empirical discriminative power and local smoothness of the feature transformation. Yi *et al.* [11] adopted canonical correlation analysis (CCA) to find two correlated spaces for NIR and VIS images. Lei and Li [12] proposed a coupled spectral linear and kernel regression model by minimizing both least square errors in two regression models and the difference between two regression coefficients. Its follow-up work in [15] refined the framework by considering both modalities in reconstruction and incorporating a Fisher linear discriminant analysis (LDA)-like objective function. Klare and Jain [5] proposed to represent the image by its kernel similarity with prototypes in this modality by assuming that a face in different modalities should keep similar relations to those prototypes in each modality. Random patches-based LDA was proposed to further improve the performance of dimensionality reduction. In [17], labels were explicitly utilized in regularized objective function to guarantee a good low-dimension representation as well as large margins between different classes. The formulated problem could be efficiently solved by following the existing fast least square solvers.

However, subspace learning methods discussed above spontaneously suffer from insufficient training sample problem, i.e., weak variations coverage, and nonoverlap identity between training and testing data. Chen *et al.* [14] utilized local neighborhood relations [40] embedded in one modality space to synthesize the same image in another modality. Similar thought has been developed through sparse coding in [18] to synthesize faces from one modality to another. Shao *et al.* [13] enhanced the quality of surveillance NIR images by superresolution from NIR to VIS and, therefore, rendered an improved high-resolution VIS image. Nevertheless, this kind of approach is vulnerable to arbitrary tests especially when there are few training samples as synthesis basis. In addition, fine alignment and postprocessing are necessary to remove blurring introduced by image synthesis. As a coding scheme, the proposed H^{2}words work at low level, and can still work well when the training and test data identities are nonoverlap. In addition, identities, in other words, multiple images for each subject required by many discriminant approach, such as [11], [12], [15], and [17], are not necessary for H^{2}words, because it only needs multimodality image pairs as well as pose and expression category information.

Spectral clustering (SC) is a principal way of using graph for data clustering [41]–[43]. There are a few variations of SC, depending on the formulations of graph Laplacian used in SC: RatioCut [44], Ncut [41], and MinMaxCut [43]. The general form of graph, hypergraph, whose hyperedges can connect more than two vertices is an ideal tool for the higher order relations analysis [45]. Recently, hypergraph and its partition have been adopted in vision problems, e.g., video segmentation [46] and image retrieval [47]. Different from them, the proposed Hwords aim to link between-modality visual words as well as within-modality ones for cross-modality feature learning. Second, the proposed weighted chi-square metric through PM improves the discriminative power of Hwords, which has not been discussed before.

PM [32] provides partial correspondence for local features in image classification. It is able to reach an approximate global geometric correspondence by spatial heuristic [33], and its kernelization works fairly well with classifiers, e.g., support vector machine (SVM). It should be noted that the way we utilize pyramid histograms is different from others as we use the difference of matching results between different levels to guide the chi-square metric. Our pyramid histograms are generated by semantically combining Hwords, while in [33], they are generated by spatial expansion of scanning windows. Although we also divide the image into patches, we repeat the histograms match in each patch and gain the pyramid by changing the resolution of histograms in each patch. Note that our work is different from [48], in which they essentially changed the resolution of images instead of histograms. Our work is also different from weighted Weber distance metric for texture segmentation [49] in that our weights work on the histograms of different resolutions rather than on different bins.

This paper is an extension of [50]. In this paper, we propose a generic Hword model that is able incorporate both handcraft codebooks (e.g., LBP [9]) and learned codebooks (e.g., LE [24]), and add more experiments to demonstrate the effectiveness of the proposed method in different scenarios. Specifically, we compare with three closely related approaches published in [5], [6], and [17], and add a new multimodality face database [51]. In addition, we show more figures/tables under different experiment settings (Figs. 11 and 12 and Tables IV–VII) to demonstrate that our method can work well in more general cases. Finally, more technical details (Tables I and II and Figs. 3–5) including proofs

TABLE I
NOTATIONS AND DESCRIPTIONS

Variable	Description
w	a visual-word
$\mathcal{V}$	a visual codebook including C words
$\mathcal{H}^{(i)}$	a hyperlingual-word indexed by i
$\mathcal{H}$	hyperlingual-words codebook
M_1, M_2	numbers of modalities and pose&expression variations
m_1, m_2	indices of modalities and pose&expression variations
M	total number of visual-words
N	total number of training data
n,m	indices of N and M, respectively
$\mathbf{t}$	histogram of a codebook
H	matrix for a hypergraph
C	number of visual-words in the initial codebook
c	index of a visual-word.
k	number of nearest neighbors of a hypergraph.
$\mathcal{T}$	visual word-context table.
$t_n^{(m)}$	a single entry of $\mathcal{T}$ indexed by n and m.
r	number of hyperlingual-words in the codebook.
$\mathcal{S}(\cdot,\cdot)$	Gaussian kernel similarity.

TABLE II
ABBREVIATIONS AND DESCRIPTIONS

Abbreviation	Description
LDA [21]	Fisher linear discriminant analysis
LDA+CCA [11]	LDA followed by canonical correlation analysis
LE [24]	learning based facial descriptor
LBP [9]	DoG filters + local binary pattern
MPL [14]	mapping learning based face synthesis
LCSR(KCSR) [12]	linear (kernel) coupled spectral regression
KPS [5]	kernel prototype similarity based recognition
CITP [6]	coupled information-theoretic encoding
LDSR(KDSR) [17]	regularized discriminative spectral regression
LE_1, LE_2, LE_3	LE descriptors with different sampling patterns
LBP_1, LBP_2, LBP_3	LBP descriptors using different sampling patterns
Hword or H(·)	proposed hyperlingual-words
H^2word or $\text{H}^2(\cdot)$	proposed hierarchical hyperlingual-words
H(LE/LBP)	Hwords with LE/LBP descriptors
H^2(LE/LBP)	H^2words with LE/LBP descriptors

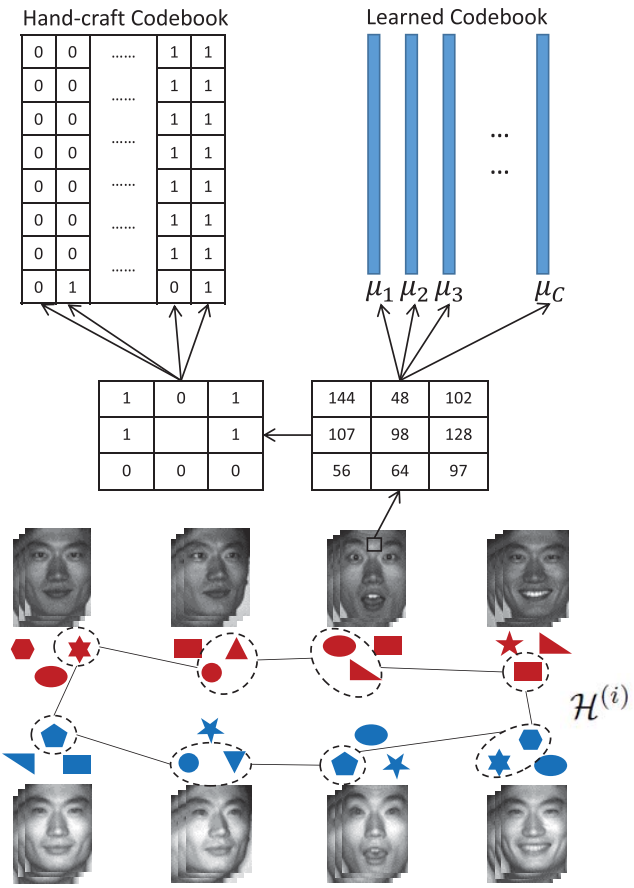

Fig. 3. Illustration of generic Hwords. Generic Hwords are able to incorporate both handcraft and learned codebooks. Each Hword, e.g., $\mathcal{H}^{(i)}$, is essentially a set of the low-level visual words from different codebooks that semantically indicates identical or similar things. Note that different colors mean visual words from different modalities, and $\mu_1, \ldots, \mu_C$ are cluster centers learned by LE.

for the key theorem (the Appendix) are provided in this extension.

III. GENERIC HYPERLINGUAL-WORDS MODELING

In this section, we will detail the motivation and the procedure of generic Hwords. For a better understanding, we summarize the variables frequently used in this section in Table I.

A. Weakly Supervised Cross-Modality Feature Learning

In conventional supervised cross-modality feature learning, there is a pair of training sets from different modalities. Taking NIR and VIS images for example, we have a typical training pair for the same identity: $\{x^{\text{nir}}_{(i,1)}, \ldots, x^{\text{nir}}_{(i,n_i^{\text{nir}})}\}$ and $\{x^{\text{vis}}_{(i,1)}, \ldots, x^{\text{vis}}_{(i,n_i^{\text{vis}})}\}$, where x is a facial feature vector, and n_i^{nir} and n_i^{vis} are the numbers of NIR and VIS images of the ith person, respectively. Then, in the test stage, given gallery images from one modality, we identify the probe images from another modality.

In this paper, we focus on a different and more practical problem, called weakly supervised cross-modality feature learning. The meaning of weakly supervised is twofold: 1) weak identity information and 2) no identity overlap between the training and test data. First, different from the supervised case with known identity, in the training stage, only an NIR-VIS image pair is given, without identity information, namely, $\{(x_i^{\text{nir}}, x_i^{\text{vis}})|1 \leq i \leq N\}$. In fact, this is very similar to the real-world applications, e.g., surveillance, where NIR-VIS image pairs are easy to collect, but identity needs more postprocessing. In addition, since we explicitly model the factors, such as facial expression and pose, they will be integrated into the Hwords learning. In the test stage, given gallery images from one modality, we still identify the probe images from another modality. Second, unlike most supervised cases, there is no identity overlap between the training and test data, which poses an even challenging problem.

B. Generic Hyperlingual-Words

Low-level feature followed by vector quantization for efficient coding has been prevalent in computer vision community recently [23], [24], [26], [27], [52]. The most important step in this line is codebook construction. The codebook can be: 1) defined by handcraft patterns [23], [26] and 2) learned by unsupervised/supervised [24], [27], [52] algorithms.

Handcraft patterns-based methods count the fixed patterns appeared in the local area and, therefore, yield a global histogram by the concatenation of these counts from each local patch. The most popular method in face recognition is LBP [23], where patterns are defined by comparing values between current pixel and its neighborhood. On the other hand, learning-based methods do not have predefined patterns. Instead, it learns the codebook by specific objective functions, namely, minimizing: 1) the total reconstruction error [53] or 2) expected distortion [27]. Then, each feature is hard- or soft-quantized to a codeword, by either nearest-neighbor rule or weighted scheme. In the Sections IV–VI, we focus on learning-based descriptor (LE) [24], since it attracts lots of research attention recently and achieves appealing performance in benchmark tests.

In this paper, we propose a generic Hwords model that is able to incorporate both handcraft patterns (e.g., LBP) and learning (e.g., LE)-based codebooks construction approaches. As shown in Fig. 3, the codebook can be built by either handcraft patterns, e.g., an 8-bit binary code in LBP, or a learned cluster center in LE, both of which are the building blocks for the proposed generic Hwords.

Suppose we have already obtained a set of codebooks $\{\mathcal{V}^{(m_1,m_2)}|1 \le m_1 \le M_1, 1 \le m_2 \le M_2\}$, where $\mathcal{V}$ is a codebook, and m_1 and m_2 index the modality and within-class variations for these codebooks. Although they are substantially different in the low level, from a high-level perspective, these words are semantically connected. For example, each codebook has some words to describe the same characteristics of the face, but from different modalities or pose/expression. If we explicitly bond these words and use them as a single word (Hword), then the new feature will tolerate both modality and pose/expression variations.

To that end, we need to first find the semantical relations between them. Then, we explicitly group visual words across both modalities and other factors, and take full advantage of hypergraph to represent this higher order relations. Finally, a few highly semantically related words will collapse to a single Hword with the following formulation:

$$\mathcal{H}^{(i)} = \{w_c^{(m_1,m_2)}|1 \le c \le C, m_1 \in [1, M_1], m_2 \in [1, M_2]\}$$

where $w_c^{(m_1,m_2)} \in \mathcal{V}^{(m_1,m_2)}$ is a single visual word. Then, we can build a set of Hwords

$$\mathcal{H} = \{\mathcal{H}^{(1)}, \mathcal{H}^{(2)} \ldots, \mathcal{H}^{(r)}\} \tag{1}$$

where r is the number of Hwords. At high-level, one can see that the Hwords are transparent across both modalities and pose and expression variations as they link low-level features by semantics. This process is shown in Fig. 3.

IV. Generic Hyperlingual-Words Learning

In this section, we discuss how to group the visual words into generic Hwords through hypergraph partition. This includes two steps: 1) graph construction and 2) partition.

A. Word-Context Table-Based Similarity

To explore the semantics between visual words and then build Hwords, we have to find out the relations between different visual words. However, finding such semantics is nontrivial, since the side information between the visual words is purely data driven. Inspired by this, we develop a data-driven approach that leverages word-context table $\mathcal{T}$ to model the high-level relations among visual words.

Suppose there are N training images in total from M_1 modalities and M_2 pose/expression variations, the word-context table $\mathcal{T}$ can be formulated as

$$\mathcal{T} = \begin{bmatrix} \mathbf{t}_1^{(1,1)} & \mathbf{t}_1^{(1,2)} & \ldots & \mathbf{t}_1^{(M_1,M_2-1)} & \mathbf{t}_1^{(M_1,M_2)} \\ \mathbf{t}_2^{(1,1)} & \mathbf{t}_2^{(1,2)} & \ldots & \mathbf{t}_2^{(M_1,M_2-1)} & \mathbf{t}_2^{(M_1,M_2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{t}_{N-1}^{(1,1)} & \mathbf{t}_{N-1}^{(1,2)} & \ldots & \mathbf{t}_{N-1}^{(M_1,M_2-1)} & \mathbf{t}_{N-1}^{(M_1,M_2)} \\ \mathbf{t}_N^{(1,1)} & \mathbf{t}_N^{(1,2)} & \ldots & \mathbf{t}_N^{(M_1,M_2-1)} & \mathbf{t}_N^{(M_1,M_2)} \end{bmatrix} \tag{2}$$

where $\mathbf{t}$ is a $1 \times C$ row vector, representing the histogram of each codebook after the vector quantization of low-level features, and $\mathcal{T}$'s columns represent different visual words, while rows represent different samples. Assume each codebook has C visual words, and let $M = C \times M_1 \times M_2$, then $\mathcal{T} \in \mathbb{R}^{N \times M}$ essentially describes the visual words' syntagmatic similarity by counting their frequencies in the training data.

To better represent the high-level relations between different words, we replace the original entries in $\mathcal{T}$ by pointwise mutual information (PMI) [54] between each training sample and each visual word. Suppose $t_n^{(m)}$ is a single entry in matrix $\mathcal{T}$, where $1 \le n \le N$ and $1 \le m \le M$, and the rowwise concatenated histogram in $\mathcal{T}$ has been normalized, then the PMI between the nth training data and the mth visual word can be expressed as follows:

$$\mathcal{T}(n,m) = \log\left(\frac{t_n^{(m)}}{\sum_n t_n^{(m)} \sum_m t_n^{(m)}}\right)$$

which constitutes the new representation of word-context table $\mathcal{T}$. Each visual word now is fully supported by $\mathcal{T}$'s column space, and the high-level similarity between two visual words w_i and w_j in $\mathcal{T}$'s ith and jth columns is formulated as $\mathcal{S}(\mathcal{T}(:,i), \mathcal{T}(:,j))$. Here, we adopt Gaussian kernel to compute this data-driven similarity with the formula of: $\mathcal{S}(\mathcal{T}(:,i), \mathcal{T}(:,j)) = \exp((- \parallel \mathcal{T}(:,i) - \mathcal{T}(:,j) \parallel_2^2)/2\sigma^2)$, where σ is the bandwidth for the Gaussian kernel, and $\mathcal{S}$ is the similarity metric.

In Fig. 3, we propose to bond visual words across both modalities and pose and expression variations. This means we strongly encourage connections between visual words from different codebooks while isolate visual words from the same codebook. Following this thought, we are able to construct a k-nearest-neighbor graph H with the following formulations:

$$H(i,j) = \begin{cases} 0 & \text{if } w_i \text{ and } w_j \in \mathcal{V}^{(m_1,m_2)} \\ \mathcal{S}(\mathcal{T}(:,i), \mathcal{T}(:,j)) & \text{otherwise} \end{cases}$$

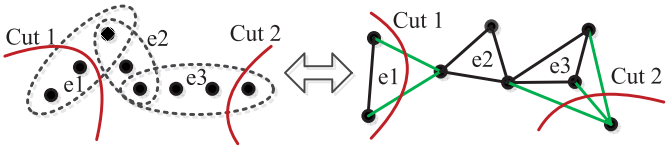

Fig. 4. Illustration of hypergraph cut. A cut on the hyperedge (left) will incur several cuts on the equivalent pairwise graph (right). The green edges in the pairwise graph are cut off in this hypergraph cut.

where $\mathcal{T}(:, j)$ should be within $\mathcal{T}(:, i)$'s k-nearest neighbors. Although such pairwise data-driven similarities are able to group visual words, they cannot explicitly handle some hidden connections preferred by Hwords. For example, if w_1 connects to both w_2 and w_3, then w_2 and w_3 should also connect, where a clique is formed among three words. To explicitly build these hidden connections, we propose to use hypergraph instead of pairwise graph, to model the correlations among words, where a hyperedge carrying the higher order relations between a word and its neighbors. Next, we will introduce how to adopt hypergraph and its partition to generate Hwords.

B. Partition for Hyperlingual-Words

In a hypergraph $G = (V, E)$, V denotes a finite set of vertices and E is a group of subsets of V. Different from the edge in a pairwise graph, hyperedge $e \in E$ can contain more than two vertices in V. All the vertices in the same hyperedge are fully connected as a clique. By assigning weight $w(e)$ to each e, we will obtain a weighted hypergraph $G(V, E, w)$. The degree of vertex $v \in e$ is defined as $d(v) = \sum_{v \in e} w(e)$, while the degree of hyperedge e is defined as $\delta(e) = |e|$, where $|\cdot|$ denotes the cardinality of this hyperedge. In addition, we use D_v, D_e, and W to indicate the diagonal matrices containing vertex degrees, hyperedge degrees, and hyperedge weights, respectively. Hypergraph G can be represented by a $|V| \times |E|$ matrix H and entry $H(v, e) = 1$ if $v \in e$ and 0 otherwise. However, binarized entries in hypergraph G may lead to information loss. To address this, we introduce the probabilistic hypergraph [47], where each e is represented by a single vertex $v_e \in V$. In this way, H can be rewritten in $H(v, e) = \mathcal{S}(v, v_e)$, where $\mathcal{S}(\cdot, \cdot)$ is a Gaussian kernel and v_e is the representative vertex for e. Considering hypergraph weight matrix and degree matrix, we can naturally formulate the hypergraph adjacency matrix as: $\Theta = HWD_e^{-1}H^T$.

To group visual words, we need to cut the hypergraph based on vertices' similarities, i.e., adjacent matrix Θ. An arbitrary cut on hypergraph $G(V, E)$ splits V into two subsets, A and its complement set A^c, such that $A \cup A^c = V$ [45], [55]. In the meanwhile, the cut hyperedge e should satisfy both $e \cap A \neq \emptyset$ and $e \cap A^c \neq \emptyset$. Suppose the hyperedge boundary $\Omega(A)$ of A is the hyperedge set that are cut, which is defined as: $\Omega(A) = \{e \in E | e \cap A \neq \emptyset \text{ and } e \cap A^c \neq \emptyset\}$, the cost of hypergraph cut is

$$\text{cut}(A, A^c) = \sum_{e \in \Omega(A)} w(e) \frac{|e \cap A||e \cap A^c|}{\delta(e)}. \tag{3}$$

The cost given in (3) can be understood in this way. Suppose each hyperedge is a fully connected clique in the pairwise graph, a cut in the hyperedge e will incur $|e \cap A||e \cap A^c|$ cuts in the clique, and the weight $w(e)$ of hyperedge e is, therefore, reweighted by $w(e)/\delta(e)$. Fig. 4 shows the hypergraph cut process along with the illustration of its counterpart: pairwise graph cut.

Similar to the pairwise graph cut, minimizing the objective in (3) over A will lead to unbalanced clusters, because the size of each cluster is not considered during the optimization. For that reason, the volume of a vertex set vol is introduced to normalize the partitions of a hypergraph, which is defined as the sum of vertices' degree. Consequently, we can write down the objective function of normalized hypergraph cut as

$$\min_{A \in V} \{\text{Ncut}(A, A^c)\} = \min_{A \in V} \left\{ \frac{\text{cut}(A, A^c)}{\text{vol}(A)} + \frac{\text{cut}(A, A^c)}{\text{vol}(A^c)} \right\}.$$

The combinatorial optimization problem above is NP-complete, and hard to solve directly. Fortunately, we are able to convert it into a real-valued eigendecomposition problem, whose eigenvectors indicate the partitions. Denote hypergraph Laplacian as $L = D_v - \Theta$, where $\Theta = \{\theta_{i,j}\}_{1 \leq i,j \leq M}$ and M is the number of vertices (total number of visual words), then L satisfies the following property.

Lemma 1: For every vector $y \in \mathbb{R}^M$, we have

$$y^T L y = \frac{1}{2} \sum_{i,j=1}^{M} \theta_{ij} (y_i - y_j)^2 \tag{4}$$

which can be trivially proved by expanding both sides. Through Lemma 1, we are able to reveal the relation between $y^T L y$ and $\text{Ncut}(A, A^c)$. Let y in Lemma 1 be the cluster indicator vector, and y_i and v_i correspond to the ith element in y and vertex set V. Each y_i can be defined as

$$y_i = \begin{cases} \sqrt{\text{vol}(A^c)/\text{vol}(A)} & \text{if } v_i \in A \\ -\sqrt{\text{vol}(A)/\text{vol}(A^c)} & \text{if } v_i \in A^c. \end{cases} \tag{5}$$

With y well-defined, we can see in the next theory that $\text{Ncut}(A, A^c)$ is equal to $y^T L y$ up to a constant factor.

Theorem 1: If y follows (5), then following conclusions hold: 1) $y^T L y = \text{vol}(V)\text{Ncut}(A, A^c)$; 2) $(D_v y)^T \mathbb{1} = 0$; and 3) $y^T D_v y = \text{vol}(V)$.

The proof of Theorem 1 is straightforward with Lemma 1 and can be found in Appendix. By Theorem 1, if we relax y to be arbitrary real values, the former normalized hypergraph cut is equivalent to

$$\min_{y \in \mathbb{R}^M} \{y^T L y\} \quad \text{s.t. } D_v y \perp \mathbb{1}, \; y^T D_v y = \text{vol}(V) \tag{6}$$

where $\mathbb{1}$ is a vector with every entry being 1. Furthermore, we substitute z for y with $z = D_v^{(1/2)} y$, and this problem becomes

$$\min_{z \in \mathbb{R}^M} \left\{ z^T D_v^{-\frac{1}{2}} L D_v^{-\frac{1}{2}} z \right\} \quad \text{s.t. } z \perp D_v^{\frac{1}{2}} \mathbb{1}, \; \|z\|_2^2 = \text{vol}(V). \tag{7}$$

The optimization problem above can be immediately solved by Rayleigh–Ritz theory, and z is given by the second eigenvector of $D_v^{-(1/2)} L D_v^{-(1/2)}$. Note that $D_v^{-(1/2)} L D_v^{-(1/2)} = I - D_v^{-(1/2)} H W D_e^{-1} H^T D_v^{-(1/2)}$ is the hypergraph Laplacian matrix proposed in [45]. Here, slightly different from theirs, we denote $D_v^{-(1/2)} L D_v^{-(1/2)}$ as L_{sym},

Algorithm 1 Hyperlingual-Words Generation

Input: N training images, $M = C \times M_1 \times M_2$ visual words included in $M_1 \times M_2$ visual codebooks.

Output: Hyperlingual-words $\mathcal{H}$.

Steps:

1: Compute the visual words context-table in Eq. (2) by N training samples and $M_1 \times M_2$ visual codebooks.
2: Set up the probabilistic hypergraph based on word context-table and k-nearest-neighbor rule.
3: Compute hypergraph matrix H, diagonal matrix D_v, D_e, W, and adjacency matrix $\Theta = HWD_e^{-1}H^T$.
4: Compute the normalized hypergraph Laplacian $L_{\text{sym}} = D_v^{-\frac{1}{2}} L D_v^{-\frac{1}{2}}$ where $L = D_v - \Theta$.
5: Compute the eigen-decomposition of L_{sym} and take the first r eigenvectors to generate indicator matrix $\hat{Y}$.
6: Cluster the row vectors of $M \times r$ indicator matrix $\hat{Y}$ by K-means and then cluster the visual words from different codebooks in the same way. Each cluster is a hyperlingual-word $\mathcal{H}^{(i)}$ in codebook $\mathcal{H}$.

and suggestively call it normalized hypergraph Laplacian. The two-way cut above (A and A^c) can be naturally extended to r-way cut by updating the definition of y

$$y_{i,j} = \begin{cases} 1/\sqrt{\text{vol}(A_j)} & \text{if } v_i \in A_j \quad (i = 1, \ldots, M \\ 0 & \text{otherwise} \quad j = 1, \ldots, r). \end{cases} \tag{8}$$

Let Y be an $M \times r$ matrix with each column as an indicator vector. Similar with that in two-way cut, we can observe in r-way cut that $Y_j^T D_v Y_j = 1$, and $Y_j^T L Y_j = \text{cut}(A_j, A_j^c)/\text{vol}(A_j)$. Removing the column index in Y, we achieve the formulation of r-way hypergraph cut as

$$\min_{A_1, A_2, \ldots, A_r} \text{Tr}(Y^T L Y) \quad \text{s.t. } Y^T D_v Y = I \tag{9}$$

where I is a identity matrix, and $\text{Tr}(\cdot)$ is the trace of a matrix. Similarly, we relax this problem by allowing Y to take real values and replacing Y by $D_v^{-(1/2)} Z$. Finally, the r-way hypergraph cut problem turns to be

$$\min_{Z \in \mathbb{R}^{M \times r}} \text{Tr}(Z^T L_{\text{sym}} Z) \quad \text{s.t. } Z^T Z = I. \tag{10}$$

If the eigenvalues are sorted increasingly, we take the first r eigenvectors $\hat{Z} = [z_1, \ldots, z_r]$ and yield a new indicator matrix $\hat{Y} \in \mathbb{R}^{M \times r}$ by $\hat{Y} = D_v^{-(1/2)} \hat{Z}$. According to the spectral theory, the most popular way is to cluster $\hat{Y}$'s row vectors $u_1, \ldots, u_M$ with K-means into r clusters $U_1, \ldots, U_r$. Then, we can obtain Hwords codebook $\mathcal{H} = \{\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(r)}\}$ by $\mathcal{H}^{(i)} = \{w_j | u_j \in U_i\}, 1 \le i \le r$. Algorithm 1 shows the entire procedure of generating Hwords.

V. Weighted Distance Metric

A. Hierarchical Hyperlingual-Words

Hwords are able to mitigate the divergence of visual words that have similar or identical semantics from different modalities or impact factors; however, this does not guarantee that the discriminative power of the visual words is improved as well. To overcome this, we propose a multiresolution histograms-based metric for image match, by assuming that there is always existing a resolution for the histogram that yields the best performance, and any histogram with different resolutions will degrade the final performance if the histogram intersections of these two resolutions are very different. Therefore, the difference between histogram intersections can be considered as an ideal weight for the distance metric, e.g., chi-square distance.

To model the difference between multiresolution histograms, we introduce the concept of suitable resolution first. It is suggested in the work of PM that the finer the histogram is, the more accurate the match will be [32], [33]. However, this is not always true. On the one hand, too many words will ruin the discriminative property, because it will lead to very sparse histogram. One subtle geometric change of object will yield dramatic changes in histograms. On the other hand, fewer visual words will push it to another extreme—it fails to sufficiently represent all variations.

In this paper, we demonstrate that for a specific problem there is a suitable resolution and it is not necessarily the highest one. Then, the suitable resolution can guide to weight chi-square metric for better performance. To generate the pyramid histograms, we design a novel structure called H^{2}words, each level of which is composed of Hwords in a specified scale. As shown in Fig. 2 (Step 4), the finest histogram can be generated by Hwords $\mathcal{H}_1$ in the smallest scale. Then, repeating Algorithm 1 with $\mathcal{H}_1$ as an input, we can obtain Hwords $\mathcal{H}_2$ in a larger scale. In this way, we can iteratively generate a group of Hwords codebooks $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_L\}$, where L indicates the total number of levels of the hierarchical structure.

B. Weighted Chi-Square Metric

The learned H^{2}words help weight the chi-square metric in the following way. Let F_a and F_b be the features of probe and gallery images, then the histogram intersection $\mathcal{I}$ can be written as

$$\mathcal{I}_l^{(1,r)}(F_a, F_b) = \sum_{i=1}^{r} \min(\mathcal{H}_l^{(i)}(F_a), \mathcal{H}_l^{(i)}(F_b)) \tag{11}$$

where i denotes the ith bin in the histogram and there are totally r bins, $\mathcal{H}_l(\cdot)$ is the histogram generated by the lth level Hwords $\mathcal{H}_l$, the subscript of $\mathcal{I}$ indicates the level of histogram, and its superscript indicates the range of Hwords. The difference between histogram intersections of two successive levels can be formulated as

$$\partial \mathcal{I}_l^{(1,r)}(F_a, F_b) = \mathcal{I}_l^{(1,r)}(F_a, F_b) - \mathcal{I}_{l-1}^{(1,2r)}(F_a, F_b) \tag{12}$$

where $\partial \mathcal{I}_l$ is always greater than 0 if we define l to be the coarser level and $l-1$ the finer level. To better understand the difference between histogram intersections, we show a simple example here. Considering the first two bins in the $(l-1)$th level and one corresponding bin in the lth level, the specific $\partial \mathcal{I}_l^{(1,1)}(F_a, F_b)$ for these two bins is: $\mathcal{I}_l^{(1,1)}(F_a, F_b) - \mathcal{I}_{l-1}^{(1,2)}(F_a, F_b)$.

Based on the analysis above, we define the weight for each bin in different levels as

$$\mathcal{W}_l^{(i)} = \begin{cases} \left(\frac{1}{2}\right)^{|l-l_s|} \Big/ \partial \mathcal{I}_l^{(i,i)} & \text{if } l > l_s \\ 1 & \text{if } l = l_s \\ \left(\frac{1}{2}\right)^{|l-l_s|} \Big/ \partial \mathcal{I}_{l+1}^{\left(\lfloor \frac{(i+1)}{2} \rfloor, \lfloor \frac{(i+1)}{2} \rfloor\right)} & \text{if } l < l_s \end{cases} \quad (13)$$

where i is the bin's index, and l_s denotes the suitable resolution of the histogram. The formulations above suppress the weight if the level of the bin is different from the ideal resolution. In addition, it will pay less attention to the level that has a large histogram intersection difference from the coarser level. Furthermore, the ith and $(i+1)$th bins will share the weights if $l < l_s$. Integrating the weighting scheme proposed in (13), the new weighted chi-square distance metric for two histograms can be written as

$$\chi_l^2\left(\mathcal{H}_l(F_a), \mathcal{H}_l(F_b)\right) = \sum_{i=1}^{r} \mathcal{W}_l^{(i)} \frac{\left(\mathcal{H}_l^{(i)}(F_a) - \mathcal{H}_l^{(i)}(F_b)\right)^2}{\left(\mathcal{H}_l^{(i)}(F_a) + \mathcal{H}_l^{(i)}(F_b)\right)}.$$

Suppose $l = 0$ indicates the finest level of histogram, while $l = L$ indicates the coarsest one, then the chi-square distance between two multiresolution histograms generated by the proposed H^{2}words can be written as

$$\chi^2(\mathcal{H}(F_a), \mathcal{H}(F_b)) = \sum_{l=0}^{L} \chi_l^2(\mathcal{H}_l(F_a), \mathcal{H}_l(F_b)) \quad (14)$$

where $\mathcal{H}(\cdot)$ is the concatenation of a group of histograms $\{\mathcal{H}_1(\cdot), \mathcal{H}_2(\cdot), \ldots, \mathcal{H}_L(\cdot)\}$.

VI. Experiments and Results

We compare our methods with the state-of-the-art methods on three recently published multimodality face databases: 1) BUAA-VisNir database; 2) Oulu-CASIA NIR&VIS database; and 3) CASIA NIR-VIS 2.0 database, as they include both two modalities and pose/expression variations.

A. Competitive Methods and Setting

The proposed method is compared with several general face recognition methods shown in Table II: LDA [21], Gabor feature [1], LBP[1] [9], learning-based descriptor (LE) [24], as well as cross-modality face recognition algorithms: mapping learning (MPL) [14], LDA+CCA [11], linear/kernel couple spectral regression (LCSR/KCSR) [12], kernel prototype similarity (KPS) [5], coupled information-theoretic encoding (CITP) [6], and regularized discriminative spectral regression (LDSR/KDSR) [17]. Both LE and LBP are adopted as the low-level features to generate codebooks. It should be noted that some competitive methods do not target at multimodality face recognition; however, as general face recognition algorithms, they can tackle pose and expression variations well. We use Hwords or H$(\cdot)$ to denote hyperlingual words, and H^{2}words or H$^2(\cdot)$ to denote hierarchical Hwords with the weighted chi-square metric proposed in (14).

[1]Note that we use chi-square metric instead of cosine distance in LBP [9]-based test, which consistently improves the performance.

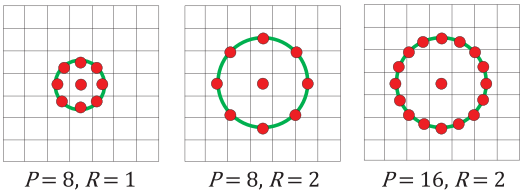

Fig. 5. Illustration of patterns used in the sampling method.

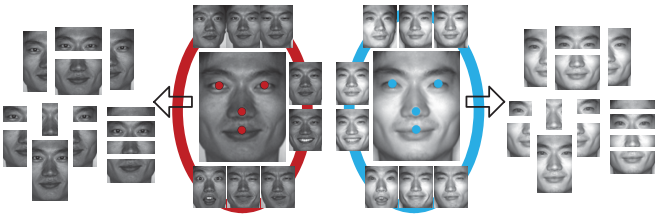

Fig. 6. Samples from the BUAA-VisNir face database. In the middle are one subject's faces with different poses or expressions in NIR and VIS. Four points on the faces are the key points used in the face partition, giving rise to 14 different components located at the left- and right-hand side of the figures.

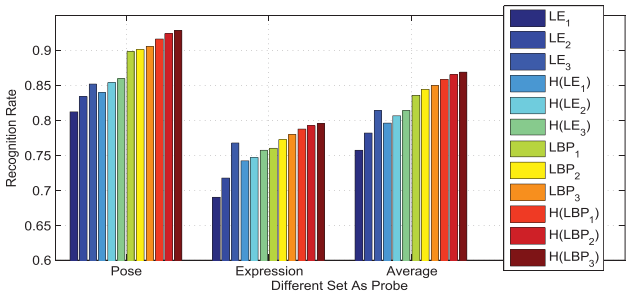

Fig. 7. Comparisons with LBP [9] and LE [24] methods. Average means average results over pose and expression. The histogram resolution for Hwords is 16.

We crop all facial images into the size of 60 × 48. For LE descriptor, we partition the whole face into components based on four key points shown in Fig. 6. Such component level feature learning has been widely discussed in face related problems, e.g., face detection [56] and face recognition [23]. On each component, the low-level features in each 6 × 6 patch are binned into histogram. Note that the LBP is implemented directly on these nonoverlap local patches for histograms. Finally, we learn specific Hwords and H^{2}words for each component, and therefore, the semantics of different components can be precisely recorded in different sets of Hwords/H^{2}words.

We use three different neighbors sampling methods for LE and LBP. Specifically, following the notation in [23], we use pair (P, R) to index the sampling methods, where P is the number of neighbors and R is the radius. In this paper, the three sampling methods are $(8, 1)$, $(8, 2)$, and $(16, 2)$, as shown in Fig. 5. In the following part, to simplify our notations, we use method$_1$ to index local descriptors using the sampling method $(8, 1)$ and method$_2$ for sampling methods $(8, 1) + (8, 2)$, and so on. We will explore the three patterns' performance in Fig. 7 by adopting them in LE, H(LE), LBP, and H(LBP).

The resolutions of pyramid histograms in H^{2}word are 8, 16, 32, and 64; and 16 is chosen as the suitable one based on fivefold cross-validation in the training set whose results can be found later in Fig. 8. For methods using principal

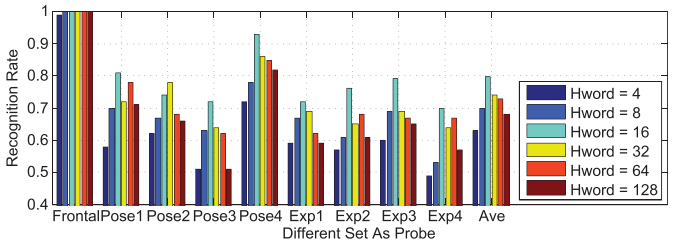

Fig. 8. Illustration of the impacts of histograms' resolutions. Note that we use H(LE$_1$) in this experiment.

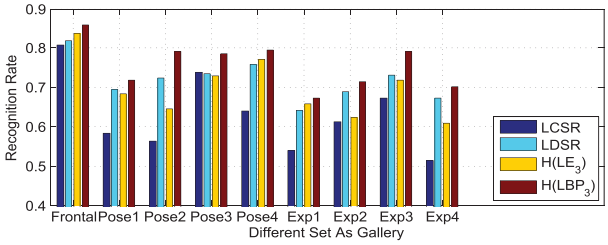

Fig. 9. Illustration of the impacts of different gallery sets.

component analysis (PCA), we choose the dimensions of the projection matrix by keeping 95% energy of corresponding eigenvalues. For methods using LDA, we use 49 dimensions, since there are totally 50 classes in the training data. In each codebook of LE, 32 visual words are produced, while in LBP, the number of predefined visual words is 59 for $(8, 1)$ and $(8, 2)$, and 243 for $(16, 2)$. We empirically set $k = 7$ for k-nearest-neighbor graph construction, and $\sigma = 10$ in computing the word-context table. For other compared methods, we tune their parameters to reach the best performance, e.g., the number of words and trees in CITP, the regularization parameters in LCSR/KCSR and LDSR/KDSR, and the number of random subspace in KPS.

B. Results on BUAA-VisNir Face Database

1) Database Descriptions: There are two parts in BUAA-VisNir face database [57], namely, NIR and VIS. There are 150 subjects included in the database, and each of them has 18 images, 9 under NIR and the other 9 under VIS. Each nine images contain nine distinct poses or expressions: 1) neutral-frontal; 2) left-rotation; 3) right-rotation; 4) tilt-up; 5) tilt-down; 6) happiness; 7) anger; 8) sorrow; and 9) surprise. Images from NIR and VIS with the same pose and expression are accurately aligned, as shown in Fig. 6. To achieve sufficient correspondence when the probe and gallery images are not in the same pose and expression, we use face components instead of holistic features. Four typical key points on faces are manually marked to guide the partition [58], i.e., center of two eyes, nose tip, and mouth center, as shown in Fig. 6.

2) Experiments Configuration: We use 900 images of 50 subjects as training data and 1800 images from the other 100 subjects as testing data. Preprocessing, e.g., histogram equalization, and DoGs are implemented before feature extraction. Note that in the testing stage, we only use one VIS image of each subject as the reference unless specified otherwise. Hence, there are 100 VIS images in the gallery and 900 NIR images in the probe. Nearest-neighbor classifier is used for all methods.

3) Results and Discussion: There are four experiments in this section: 1) the impacts of different local descriptors and sampling patterns; 2) the impacts of different resolutions of Hwords; 3) the impacts of different gallery images; and 4) the impacts of different comprehensive comparisons. First, we take a close observation on four pattern-dependent local methods, i.e., LE, LBP, H(LE), and H(LBP) in Fig. 7. We fix the gallery as the neutral-frontal VIS image and vary the test NIR images. One can see that the proposed Hwords work better

TABLE III

RECOGNITION RESULTS ON BUAA-VISNIR FACE DATABASE. NOTE THAT THE METHODS IN THE FIRST GROUP ARE GENERAL FACE RECOGNITION METHODS, WHILE THOSE IN THE SECOND GROUP ARE SPECIFICALLY DESIGNED FOR MULTIMODALITY FACE RECOGNITION

Method	Poses	Expressions	Average
LDA [21]	0.450	0.405	0.430
Gabor [1]	0.628	0.520	0.580
LE$_3$ [24]	0.852	0.768	0.814
LDA+CCA [11]	0.692	0.525	0.618
MPL$_3$ [14]	0.602	0.445	0.532
CITP [6]	0.642	0.563	0.608
KPS [5]	0.702	0.62	0.666
LCSR [12]	0.884	0.713	0.808
KCSR [12]	0.872	0.742	0.814
LDSR [17]	0.858	0.755	0.812
KDSR [17]	0.878	0.770	0.830
LBP$_3$ [9]	0.906	0.780	0.850
H(LE$_3$) (Ours)	0.860	0.758	0.814
H^2(LE$_3$) (Ours)	0.898	0.796	0.853
H(LBP$_3$) (Ours)	0.929	0.795	0.869
H^2(LBP$_3$) (Ours)	0.948	0.813	0.888

on pose or expression variations than original LE and LBP, except for LE$_3$. We believe that more sampling patterns could provide discriminative features for LE and dominate the results. In addition, the more sampling patterns are used, the better the performance is. Second, we experimentally prove the existence of the suitable resolution for histograms. In Fig. 8, Hword $= 16$ works better than other resolutions in the fivefold cross-validation in the training set. Therefore, it is rational to empirically use this suitable resolution for the weighted chi-square metric in the following experiments when image resolution is fixed.

In addition to demonstrating Hwords can connect visual words with semantics, we use faces with different poses and expressions as the gallery images rather than only using the frontal neutral one, which is shown in Fig. 9. No matter which pose or expression we take as the gallery, our method performs better in most cases, especially for H(LBP$_3$). It further proves that the proposed Hwords can capture high-level semantics ignored by low-level features. We further detail the improvement by Hwords in Table IV by using different probe and gallery images. Detailed increase or decrease is indicated by marks $\uparrow$ or $\downarrow$ and numbers afterward. In most cells, the accuracy is enhanced (up to 0.11).

Extensive results are shown in Tables III and V and Fig. 10, from which we can conclude our method is superior to other state-of-the-art methods. Similarly, we use one frontal-neutral VIS image as the gallery for each subject and all the other NIR images as the probe. Since there is no overlap identities

TABLE IV

COMPARISONS OF LBP_3 AND $H(LBP_3)$ ON BUAA-VisNir DATABASE BY VARYING TEST (COLUMN) AND GALLERY (ROW) IMAGE SETS OVER NINE DIFFERENT POSES/EXPRESSIONS. THE FIRST NUMBER IN EACH CELL IS THE RESULT FROM LBP, WHILE THE SECOND ONE SHOWS EITHER INCREASE ↑ OR DECREASE ↓ BY $H(LBP_3)$

Pose/Exp	Frontal	Pose1	Pose2	Pose3	Pose4	Exp1	Exp2	Exp3	Exp4
Frontal	1↔	0.85 ↑.02	0.86↑.06	0.85↑.03	0.92↓.01	0.72↑.01	0.76↑.06	0.79↑.03	0.77↔
Pose1	0.79↑.02	1↔	0.7↑.02	0.7↑.07	0.74↑.03	0.49↑.04	0.61↓.01	0.65↑.07	0.61↓.06
Pose2	0.92↔	0.69↑.05	1↔	0.77↑.01	0.9↓.02	0.56↑.07	0.7↑.03	0.8↑.02	0.63↓.01
Pose3	0.88↑.04	0.74↑.01	0.74↑.08	1↔	0.7↑.01	0.48↑.10	0.82↓.01	0.76↑.07	0.59↑.06
Pose4	0.94↑.01	0.81↔	0.83↔	0.71↑.05	1↔	0.67↑.01	0.67↔	0.71↔	0.74↔
Exp1	0.66↑.08	0.39↑.03	0.53↑.04	0.56↑.02	0.56↑.07	1↔	0.70↓.03	0.71↑.06	0.56↑.11
Exp2	0.75↓.01	0.57↔	0.59↓.04	0.76↔	0.56↑.06	0.54↑.06	1↔	0.78↑.05	0.63↑.06
Exp3	0.80↑.03	0.65↔	0.71↑.05	0.80↑.02	0.69↓.02	0.71↑.04	0.84↓.01	0.99↔	0.74↑.08
Exp4	0.72↑.04	0.53↑.03	0.56↑.07	0.56↑.04	0.60↑.06	0.53↑.10	0.71↑.01	0.71↑.04	1↔

TABLE V

RANK-1 AND VERIFICATION ACCURACY ON BUAA-VisNir FACE DATABASE

Method	Rank-1	FP = 0.1%	FP = 1%
MPL_3 [14]	0.532	0.333	0.581
LDA+CCA [11]	0.618	0.427	0.690
KPS [5]	0.666	0.417	0.602
LCSR [12]	0.808	0.630	0.846
KCSR [12]	0.814	0.667	0.838
LDSR [17]	0.812	0.674	0.813
KDSR [17]	0.830	0.695	0.868
$H^2(LE_3)$ (Ours)	0.853	0.699	0.847
$H^2(LBP_3)$ (Ours)	0.888	0.734	0.888

between the training and test sets, not surprisingly, we see that the local features work better than holistic ones, e.g., LDA and LDA + CCA, which need sufficient training samples to accurately model the subspace. Moreover, for LE and LBP, the proposed hierarchical structure further enhances the performance, compared with their Hwords version. This proves that the weighted metric is helpful in practice. We note that MPL does not work as well as in [14]. This is because MPL is training data sensitive and needs sufficient prototype images to synthesize the new one under the target environment. Other methods, such as KPS, CITP, LCSR/KCSR, and LDSR/KDSR, are affected by nonoverlap training and test data, lack of identity information, or do not explicitly consider the variations caused by pose and expression. However, our method still performs well in weakly supervised case, especially when we shuffle the gallery data (Table IV).

C. Results on Oulu-CASIA NIR&VIS Database

1) Database Descriptions: We adopt a subset of Oulu-CASIA NIR&VIS database [14] for our second experiment. Sample faces can be found in Fig. 1. The entire database contains 80 subjects (50 from Oulu University and 30 from CASIA), each of which comprises six expressions, i.e., anger, disgust, fear, happiness, sadness, and surprise. We select 40 subjects from this database, namely, 10 from Oulu University and 30 from CASIA, and randomly select 8 images from each of six expressions from both NIR and VIS to build a data set for evaluation. Therefore, there are 96 (48 NIR images and 48 VIS images) images in total for each subject.[2]

[2]Note that this is different from our previous setting in [50]. The new setting here aims to mitigate the divergence between two parts of Oulu-CASIA NIR&VIS database.

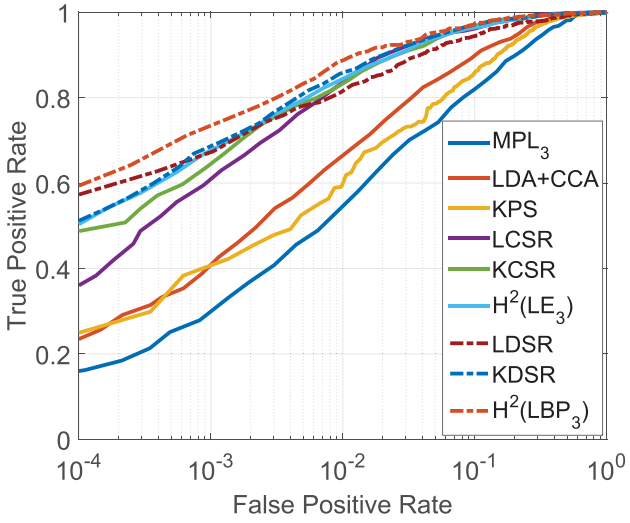

Fig. 10. ROC curves compared with other existing methods on BUAA-VisNir face database. We rescale the x-axis for a better illustration.

TABLE VI

RANK-1 AND VERIFICATION ACCURACY ON Oulu-CASIA NIR&VIS FACE DATABASE

Method	Rank-1	FP = 0.1%	FP = 1%
MPL_3 [14]	0.489	0.114	0.419
LDA+CCA [11]	0.589	0.200	0.453
KPS [5]	0.622	0.222	0.483
LCSR [12]	0.653	0.225	0.488
KCSR [12]	0.660	0.261	0.497
LDSR [17]	0.686	0.300	0.601
KDSR [17]	0.669	0.319	0.561
$H^2(LE_3)$ (Ours)	0.670	0.261	0.503
$H^2(LBP_3)$ (Ours)	0.708	0.336	0.620

2) Experiments Configuration: Half of the data set, 20 subjects' images are used as training and the others as testing sets. All the parameter configurations are the same as those in BUAA-VisNir database. We report recognition accuracy, receiver operating characteristic (ROC) curves, and true positive rates given false positive rate in Table VI and Fig. 11. Note that in the test stage, slightly different from the last experiment, we use all VIS images (instead of single frontal VIS images) of 20 testing subjects as the gallery and all corresponding NIR images as the probe.

3) Results and Discussion: As we can see, the average performances of these methods are poor compared with those on BUAA-VisNir, though the number of people is decreased to 40. There are several reasons. First, NIR and

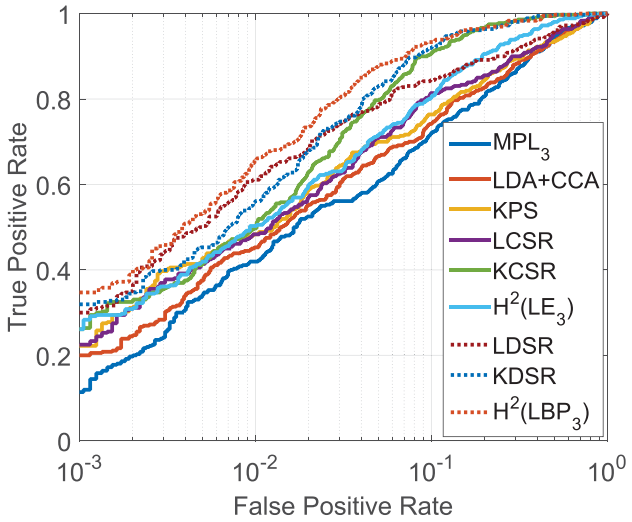

Fig. 11. ROC curves compared with other existing methods on Oulu-CASIA NIR&VIS database. We rescale the x-axis for a better illustration.

its corresponding VIS image are not well-aligned in this database, since these two sorts of images are not captured at the same time without any pose or expression changes. This gives great impacts on almost all heterogeneous feature learning methods that need good feature alignment between different modalities. Second, Oulu-CASIA NIR&VIS database consists of two parts (one from Oulu U and another from CASIA), and they are captured in two slightly different environments, so the illumination conditions are different, especially for VIS images. This largely affects the performance of MPL, which assumes the consistence between the training and testing data when reconstructing images. Finally, unlike BUAA-VisNir database, the expressions in Oulu-CASIA NIR&VIS database are relatively exaggerated, giving rise to dramatic nonrigid deformation on faces. Though there are many factors degrading the performance of all methods, our method is still comparable with most of them and sometimes outperforms others in this extremely challenging test.

D. Results on CASIA NIR-VIS 2.0

1) Database Descriptions: CASIA NIR-VIS is another multimodality face database, including both NIR and VIS images [51]. There are four sessions and in total 725 subjects in the database, and the identities of each session may or may not overlap other sessions. The database includes people from different aging groups, in different poses, expressions, or accessories. For each subject, there are 1–22 VIS images and 5–50 NIR images.

2) Experiments Configuration: In this paper, we take session 2 for our evaluations, which has 308 different subjects. Images from the first 200 subjects are adopted as the training samples and the rest are the test samples. For each subject in the training set, eight NIR images are selected for evaluations to cover as much pose or expression variations as possible. Then, we choose their nearest neighbors from VIS modality, and obtain another eight VIS images. For each subject in the test set, eight NIR images and four VIS images are selected as probe and gallery images, respectively. Following the weakly supervised cross-modality feature learning setting, only

TABLE VII

RANK-1 AND VERIFICATION ACCURACY ON CASIA NIR-VIS 2.0 FACE DATABASE

Method	Rank-1	FP = 0.1%	FP = 1%
MPL$_3$ [14]	-	-	-
LDA+CCA [11]	0.258	0.039	0.186
KPS [5]	0.282	0.037	0.174
LCSR [12]	0.348	0.077	0.300
KCSR [12]	0.338	0.076	0.285
LDSR [17]	0.362	0.088	0.316
KDSR [17]	0.375	0.093	0.330
H^2(LE$_3$) (Ours)	0.351	0.082	0.298
H^2(LBP$_3$) (Ours)	0.438	0.101	0.365

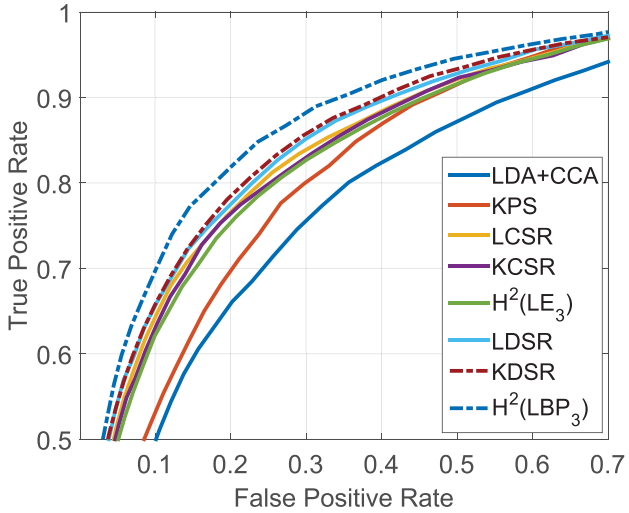

Fig. 12. ROC curves compared with other existing methods on CASIA NIR-VIS 2.0 database.

NIR-VIS pairs are given in the training stage, without identity information. To obtain the within-modality variations information, we use K-means to partition NIR and VIS training images into four groups, respectively, and thus generate four pseudo within-modality variations. Then, $4 \times 2 = 8$ codebooks can be learned accordingly. Note that we do not evaluate the MPL on this database, since there are no aligned NIR-VIS pairs.

3) Results and Discussion: As we can see from Table VII and Fig. 12, H^{2}words with LE as the local descriptor perform comparably with other methods, while that with LBP descriptor performs better than others. Notably, the overall performance is relatively low, since images from both modalities are subject to pose, expression, and illumination variations. The key impact factor, we believe, is the lighting conditions of VIS images. Unlike NIR images, VIS images are not illumination-invariant, which causes a dramatic degeneration. Such problems have been widely discussed in face recognition problems before [1], [59]. Besides, due to lack of identity information, conventional supervised cross-modality feature learning methods do not perform as what we expected. In addition, since there are no ground truth labels for within-modality variations, such as pose or expression, the pseudolabels learned by K-means may also affect the system performance.

VII. Conclusion

In this paper, to tackle the challenging multimodality face recognition problem in a weakly supervised fashion, we proposed the new visual descriptors called generic Hwords to incorporate both handcraft and learned codebooks. Then, H^2words were designed to generate robust features against both cross-modality and within-modality variations. Third, by leveraging the suitable resolution of histograms, we derived a new weighted chi-square metric for classification. Finally, extensive experimental results demonstrated that our method is effective on multimodality face recognition problem and works better than the state-of-the-art methods on three multimodality face databases.

Appendix

Proof of Theorem 1:

Proof: For conclusion (1), according to Lemma 1. We have

$$\begin{aligned} f^T Lf &= \frac{1}{2}\sum_{i,j=1}^{M}\theta_{ij}(f_i - f_j)^2 \\ &= \frac{1}{2}\sum_{v_i\in A, v_j\in A^c}^{M}\theta_{ij}\left(\sqrt{\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)}}+\sqrt{\frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)}}\right)^2 \\ &\quad +\frac{1}{2}\sum_{v_i\in A^c, v_j\in A}^{M}\theta_{ij}\left(-\sqrt{\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)}}-\sqrt{\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)}}\right)^2 \\ &= \mathrm{cut}(A, A^c)\left(\frac{\mathrm{vol}(A)+\mathrm{vol}(A^c)}{\mathrm{vol}(A)}+\frac{\mathrm{vol}(A)+\mathrm{vol}(A^c)}{\mathrm{vol}(A^c)}\right) \\ &= \mathrm{Ncut}(A, A^c)\mathrm{vol}(V). \end{aligned}$$

For conclusion (2), according to the definition of D_v and f

$$\begin{aligned} (D_v f)^T\mathbb{1} &= \sum_{v_i\in A} d_i\sqrt{\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)}} - \sum_{v_i\in A^c} d_i\sqrt{\frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)}} \\ &= \mathrm{vol}(A)\sqrt{\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)}} - \mathrm{vol}(A^c)\sqrt{\frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)}} = 0. \end{aligned}$$

For conclusion (3), similar to (2) we have

$$\begin{aligned} f^T D_v f &= \sum_{v_i\in A} d_i\frac{\mathrm{vol}(A^c)}{\mathrm{vol}(A)} + \sum_{v_i\in A^c} d_i\frac{\mathrm{vol}(A)}{\mathrm{vol}(A^c)} \\ &= \mathrm{vol}(A) + \mathrm{vol}(A^c) = \mathrm{vol}(V). \end{aligned}$$

■

References

[1] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 721–732, Jul. 1997.

[2] S. Z. Li and A. K. Jain, Eds., *Handbook of Face Recognition*, 2nd ed. London, U.K.: Springer-Verlag, 2011.

[3] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 627–639, Apr. 2007.

[4] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.

[5] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

[6] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 513–520.

[7] W. Luo *et al.*, "Synthesizing oil painting surface geometry from a single photograph," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 885–892.

[8] H. Zhou, Z. Kuang, and K.-Y. K. Wong, "Markov weight fields for face sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1091–1097.

[9] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Proc. 3rd IAPR Int. Conf. Biometrics*, Alghero, Italy, Jun. 2009, pp. 209–218.

[10] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. 9th Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 13–26.

[11] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching between near infrared and visible light images," in *Proc. 2nd IAPR Int. Conf. Biometrics*, Seoul, Korea, Aug. 2007, pp. 523–530.

[12] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1123–1128.

[13] M. Shao, Y. Wang, and Y. Wang, "A super-resolution based method to synthesize visual images from near infrared," in *Proc. 16th IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 2453–2456.

[14] J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikäinen, "Learning mappings for face synthesis from near infrared to visual light images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 156–163.

[15] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1707–1716, Dec. 2012.

[16] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.

[17] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 353–362, Jan. 2013.

[18] Z. Zhang, Y. Wang, and Z. Zhang, "Face synthesis from near-infrared to visual light via sparse representation," in *Proc. Int. Joint Conf. Biometrics*, Washington, DC, USA, Oct. 2011, pp. 1–6.

[19] Z. Zhang, Y. Wang, Z. Zhang, and G. Zhang, "Face synthesis from near-infrared to visual light spectrum using quotient image and kernel-based multifactor analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Barcelona, Spain, Jul. 2011, pp. 1–4.

[20] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[21] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[22] X. He and P. Niyogi, "Locality preserving projections," in *Proc. 17th Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2003, pp. 153–160.

[23] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. 8th Eur. Conf. Comput. Vis.*, Prague, Czech Republic, May 2004, pp. 469–481.

[24] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2707–2714.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[27] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 524–531.

[28] W. J. Hutchins and H. L. Somers, *An Introduction to Machine Translation*. London, U.K.: Academic, 1992.

[29] L. Ballesteros and B. Croft, "Dictionary methods for cross-lingual information retrieval," in *Database and Expert Systems Applications*. Zürich, Switzerland: Springer, Sep. 1996, pp. 791–801.

[30] K. Kishida, "Technical issues of cross-language information retrieval: A review," *Inf. Process. Manage.*, vol. 41, no. 3, pp. 433–455, 2005.

[31] G. Grefenstette, Ed., *Cross-Language Information Retrieval*, vol. 2. New York, NY, USA: Springer, 1998.

[32] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. 10th Int. Conf. Comput. Vis.*, Beijing, China, Oct. 2005, pp. 1458–1465.

[33] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 2169–2178.
[34] S. Liu, D. Yi, Z. Lei, and S. Z. Li, "Heterogeneous face image matching using multi-scale features," in *Proc. 5th IAPR Int. Conf. Biometrics*, New Delhi, India, Mar./Apr. 2012, pp. 79–84.
[35] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
[36] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 289–302, Feb. 2014.
[37] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1701–1708.
[38] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. 28th Annu. Conf. Adv. Neural Inf. Process. Syst.*, Montréal, QC, Canada, Dec. 2014, pp. 1988–1996.
[39] M. Shao, Z. Ding, and Y. Fu, "Sparse low-rank fusion based deep features for missing modality face recognition," in *Proc. 11th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Ljubljana, Slovenia, May 2015, pp. 1–6.
[40] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, Dec. 2003.
[41] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[42] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. 15th Annu. Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2001, pp. 849–856.
[43] C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *Proc. IEEE Int. Conf. Data Mining*, San Jose, CA, USA, Nov. 2001, pp. 107–114.
[44] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 11, no. 9, pp. 1074–1085, Sep. 1992.
[45] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. 20th Annu. Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 1601–1608.
[46] Y. Huang, Q. Liu, and D. Metaxas, "Video object segmentation by hypergraph cut," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1738–1745.
[47] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3376–3383.
[48] E. Hadjidemetriou, M. D. Grossberg, and S. K. Nayar, "Multiresolution histograms and their use for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 831–847, Jul. 2004.
[49] J. Chen, G. Zhao, M. Salo, E. Rahtu, and M. Pietikäinen, "Automatic dynamic texture segmentation using local descriptors and optical flow," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 326–339, Jan. 2013.
[50] M. Shao and Y. Fu, "Hierarchical hyperlingual-words for multi-modality face classification," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Shanghai, China, Apr. 2013, pp. 1–6.
[51] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, Jun. 2013, pp. 348–353.
[52] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 543–550.
[53] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1794–1801.
[54] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 461–468.
[55] S. Agarwal, K. Branson, and S. Belongie, "Higher order learning with graphs," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 17–24.
[56] J. Chen, X. Chen, J. Yang, S. Shan, R. Wang, and W. Gao, "Optimization of a training set for more robust face detection," *Pattern Recognit.*, vol. 42, no. 11, pp. 2828–2840, 2009.
[57] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001, Jul. 2012.
[58] J.-J. Wong and S.-Y. Cho, "A face emotion tree structure representation with probabilistic recursive neural network modeling," *Neural Comput. Appl.*, vol. 19, no. 1, pp. 33–54, 2010.
[59] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

Ming Shao (S'11) received the B.E. degree in computer science, the B.S. degree in applied mathematics, and the M.E. degree in computer science and engineering from Beihang University, Beijing, China, in 2006, 2007, and 2009, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA.

He was a Research Intern with the Samsung Research Laboratory, Beijing, in 2010, the Video Analytics Research Group, Motorola Solutions, Schaumburg, IL, USA, in 2013, and Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, in 2014. His current research interests include sparse modeling, low-rank matrix analysis, and applied machine learning on social media analytics.

Dr. Shao was a recipient of the Presidential Fellowship from University at Buffalo, The State University of New York, Buffalo, NY, USA, from 2010 to 2012, and the Best Paper Award Winner of the IEEE International Conference on Data Mining Workshop on Large Scale Visual Analytics in 2011. He has served as a Reviewer for the IEEE journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, Xi'an, China, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA.

He was a Tenure-Track Assistant Professor with the Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA, from 2010 to 2012. He has been a tenured faculty member with Northeastern University, Boston, USA. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. His current research interests include machine learning, computer vision, social media analytics, and big data mining.

Dr. Fu is a Lifetime Member of the Association for Computing Machinery, the Association for the Advancement of Artificial Intelligence, the International Society for Optics and Photonics, the Optical Society, and the Institute of Mathematical Statistics, and a member of the International Neural Network Society (INNS). He was a Beckman Graduate Fellow from 2007 to 2008. He received five Young Investigator Awards, such as the 2016 IEEE Computational Intelligence Society Outstanding Early Career Award, the 2015 National Academy of Engineering U.S. Frontiers of Engineering Award, the 2014 ONR Young Investigator Award, the 2014 ARO Young Investigator Award, and the 2014 INNS Young Investigator Award, five best paper awards, such as the 2014 SIAM International Conference on Data Mining, the 2013 International Conference on Automatic Face and Gesture Recognition, the 2011 IEEE International Conference on Data Mining-LSVA, the 2010 IAPR International Conference on Frontiers in Handwriting Recognition, and the 2007 IEEE International Conference on Image Processing, two Industrial Research Awards, such as the 2015 Adobe Faculty Research Award and the 2010 Google Faculty Research Award, and two Service Awards, such as the 2012 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Associate Editor Award and the 2011 IEEE International Conference on Multimedia and Expo Best Reviewer Award. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He serves as an Associate Editor, the Chair, a PC Member, and a Reviewer of many top journals and international conferences/workshops.